Graphical Abstract

GHMSCHRFT: An AI-Based De-identification Pipeline for Dutch Clinical Reports

Luc Builtjes, Joeran S. Bosma, Judith Lefkes, Anouk Veldhuis, Jeroen Geerdink, Tijmen de Haas, Fabian R. Hoitsma, Rianne A. Weber, Sarah de Boer, Myrthe Buser, Harry J. Kendall, Delano Sanches, Bram van Ginneken, Henkjan Huisman, Ewoud J. Smit, Mathias Prokop, Alessa Hering

The GHMSCHRFT Dutch medical text de-identification pipeline

Input data

Raw reports contain PHI
Names, dates, locations, identifiers must be removed before reuse.

Medical report

Rapport: **47195**

68 jarige man werd op **12-02-2024** gezien op de afdeling radiologie. CT thorax vervaardigd op verzoek van dr. **Janssen**. Vergeleken met eerder onderzoek van **03-11-2023**.

Conclusie: Stabiel beeld.
Controle over 6 maanden.

PHI tagging

RoBERTa tags 24 PHI types
Fine-tuned on 3,941 annotated Dutch medical reports from two hospitals.

Medical report

Rapport: **<DOCUMENTID>**

<LEEFTIJD> jarige man werd op **<DATUM>** gezien op de afdeling radiologie. CT thorax vervaardigd op verzoek van dr. **<PERSOON>**. Vergeleken met eerder onderzoek van **<DATUM>**.

Conclusie: Stabiel beeld.
Controle over 6 maanden.

Hiding In Plain Sight

PHI is replaced by realistic surrogates
Structure and readability are preserved.

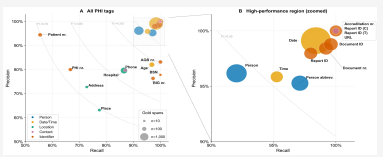
Medical report

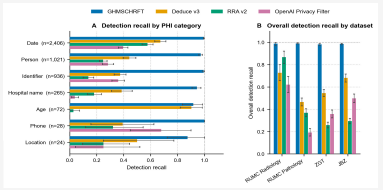
Rapport: **11246**

55 jarige man werd op **08-09-2024** gezien op de afdeling radiologie. CT thorax vervaardigd op verzoek van dr. **Vermeer**. Vergeleken met eerder onderzoek van **19-05-2023**.

Conclusie: Stabiel beeld.
Controle over 6 maanden.

Results





Detection recall: 0.985 (internal) - 0.986 (external)
F1-score: 0.943 (internal) - 0.950 (external)
Readers found: max. 3/21 leaked PHI spans

Conclusion

GHMSCHRFT substantially outperforms rule-based alternatives and generalizes to unseen radiology, pathology and emergency care reports.
High detection recall combined with realistic surrogate replacement provides strong practical privacy protection.
The detection model, Dutch HIPS module, and annotation guidelines are publicly available.

GHMSCHRFT: An AI-Based De-identification Pipeline for Dutch Clinical Reports

Luc Builtjes[a,d,*], Joeran S. Bosma[a], Judith Lefkes[b], Anouk Veldhuis[c], Jeroen Geerdink[c], Tijmen de Haas[a], Fabian R. Hoitsma[a], Rianne A. Weber[a], Sarah de Boer[a], Myrthe Buser[a], Harry J. Kendall[a], Delano Sanches[a], Bram van Ginneken[a], Henkjan Huisman[a], Ewoud J. Smit[a], Mathias Prokop[a], Alessa Hering[a]

[a]*Department of Medical Imaging, Radboudumc, Geert Grooteplein Zuid 10, Nijmegen, 6525GA, The Netherlands*
[b]*Department of Pathology, Radboudumc, Geert Grooteplein Zuid 10, Nijmegen, 6525GA, The Netherlands*
[c]*Department of Information Management, Ziekenhuisgroep Twente, Zilvermeeuw 1, Almelo, 7609PP, The Netherlands*
[d]*Department of Medical Imaging, Jeroen Bosch Ziekenhuis, Deutersestraat 2, 's-Hertogenbosch, 5223GZ, The Netherlands*

Abstract

Objective: Secondary use of clinical free-text requires reliable de-identification, but existing open-source methods do not yet perform well enough for practical deployment without institution specific tuning. We aimed to develop and validate GHMSCHRFT (from Dutch *geheimschrift*, 'cipher'), an open-source de-identification pipeline for Dutch clinical text, and to assess whether its protection holds up to human scrutiny after deployment.
Methods: GHMSCHRFT combines a transformer-based model (RoBERTa-large, pretrained on Dutch medical text) fine-tuned for PHI detection on 3,941 annotated reports from two hospitals, spanning radiology, pathology, and multiple outpatient specialties, with a Dutch-specific Hiding in Plain Sight (HIPS) surrogate replacement module. Performance was evaluated on internal (980 reports) and external (340 reports from an unseen institution) test sets, and a reader study assessed residual disclosure risk by having five readers attempt to identify leaked PHI in HIPS-processed output.

*Corresponding author
Email address: luc.builtjes@radboudumc.nl (Luc Builtjes)

Results: GHMSCHRFT achieved detection recall of 0.985 and 0.986, with precision of 0.941 and 0.949, on internal and external test sets, compared with recall of 0.549/0.682 for DEDUCE, 0.386/0.295 for an in-house rule-based system, and 0.339/0.498 for the OpenAI Privacy Filter. In the reader study, individual detection rates ranged from 0–14.3% of evaluable leaked PHI spans (5 of 21 identified collectively), with re-identification precision no higher than 8.3%.

Conclusion: GHMSCHRFT generalizes across institutions for radiology, pathology, and emergency care reports, outperforming rule-based alternatives by more than 30 percentage points in detection recall. On top of these detection results, the reader study confirms that the full pipeline provides strong practical protection: readers performed no better than chance at distinguishing genuine leaks from correctly replaced surrogates. The detection model, Dutch HIPS module, and annotation guidelines are publicly available.

Keywords: de-identification, protected health information, clinical natural language processing, hiding-in-plain-sight, GDPR

1. Introduction

Electronic health records typically contain two broad categories of data: structured fields such as diagnosis codes and date of birth, and unstructured free-text reports created by clinicians in natural language. Structured data are directly machine-interpretable, but much of the clinically relevant information is expressed in narrative text [1, 2]. Free-text reports capture longitudinal reasoning, temporal relations between findings, and contextual nuance that structured fields cannot easily represent [3]. This makes clinical narratives a valuable resource for secondary use, especially for the development of data-driven clinical tools [4].

Realizing this potential requires de-identifying clinical narratives before patient data can be shared and reused responsibly. Clinical narratives routinely contain protected health information (PHI): direct and quasi-identifiers such as person names, dates, locations, and institutional identifiers that can be used to re-identify individuals. In the European Union, the General Data Protection Regulation (GDPR) governs the processing of such personal data [5]. De-identification, the process of detecting and removing or replacing PHI to a standard that is robust against reasonably likely means of re-identification,

is therefore a prerequisite for compliant data sharing and secondary use of clinical text.

Automated de-identification has been an active research area for several decades. Early systems relied on handcrafted rules, dictionaries, and regular expressions to detect identifiers [6]. These approaches work well for structurally regular identifiers such as phone numbers or postal codes, but require substantial manual effort to develop, are fragile under format variation, and tend to generalize poorly when applied outside the institution for which they were designed. The introduction of large annotated corpora and pretrained language models shifted the field toward data-driven approaches. Transformer-based sequence-labeling models now represent the dominant paradigm and consistently outperform rule-based systems when sufficient training data are available [7, 8]. Generative large language models (LLMs) have more recently been explored as an alternative to supervised sequence labeling [9, 10]. These prompt-based approaches avoid the need for annotated training data, but in practice require careful prompt engineering and are sensitive to phrasing, output format, and model version, making reliable deployment difficult [11]. Fine-tuning generative LLMs on annotated PHI data can improve consistency, but introduces a separate privacy concern: language models are known to memorize fragments of their training data, and a publicly released fine-tuned model may regurgitate patient information seen during training [12]. This tension between performance and safe release makes generative LLMs a poor fit for de-identification pipelines intended for open distribution.

On top of this, both supervised and generative approaches have been developed and evaluated predominantly on English clinical text [13], and it remains unclear how either transfers to other languages or institutional settings. A recent example is the OpenAI Privacy Filter [14], an open-source model developed specifically for de-identification but trained on general, primarily English text, which illustrates both the progress and the remaining language gap.

For Dutch clinical text, the landscape looks markedly different; Dutch hospitals have largely developed de-identification solutions independently, but little of this work has been released openly. The main exception is DEDUCE [15], a rule-based system developed for de-identification of Dutch clinical notes, which has become the de facto open-source standard in the Netherlands. No publicly available transformer-based alternative exists for Dutch. This situation is compounded by the difficulty of sharing clinical data across institutions under GDPR, which limits the assembly of the multicenter

annotated corpora needed to train and rigorously evaluate learning-based approaches. As a result, it remains unclear whether transformer models trained on Dutch clinical data from one or more institutions can achieve the detection performance and cross-institutional generalizability required for practical deployment.

Detection performance alone, however, does not fully determine privacy risk. A common approach after detection is to mask or delete identified spans. This leaves the document structure intact but creates an unambiguous signal: any identifier that escapes detection remains in plain text while everything around it is redacted, making residual PHI easy to spot. Hiding in Plain Sight (HIPS) was proposed as an alternative strategy to address this vulnerability [16]. Rather than removing detected identifiers, HIPS replaces them with realistic surrogates, so that any residual PHI blends into a document that still reads as natural clinical text. Carrell et al. demonstrated that this approach meaningfully reduces the detectability of leaked identifiers by both automated systems and human reviewers [17]. Despite this evidence, HIPS implementations remain rare and almost exclusively English. No open-source Dutch medical module exists. Naively applying English surrogate generators to Dutch text would produce implausible output, since Dutch name distributions, address formats, and institutional identifier structures differ substantially from their English counterparts. This undermines the core idea of HIPS: residual PHI would stand out against foreign-language surrogates rather than blend in, becoming identifiable rather than protected.

Together, these gaps raise two related questions. Can a transformer model trained on multicenter Dutch clinical data achieve detection performance sufficient for practical deployment, including generalization to hospitals not seen during training? And does replacing detected identifiers with Dutch-specific surrogates reduce residual disclosure risk to a degree that is empirically verifiable through human assessment?

In this study, we address both questions by presenting and evaluating GHM-SCHRFT (from Dutch *geheimschrift*, 'cipher'), an end-to-end de-identification pipeline for Dutch clinical free-text reports. We fine-tune a Dutch clinical transformer model on multicenter annotated data from two hospitals and evaluate detection performance on internal and external test sets. To address the absence of Dutch-specific surrogate replacement tooling, we developed `dutch-med-hips` [18], an open-source HIPS module that replaces detected PHI with realistic Dutch surrogates. We assess whether the pipeline provides sufficient protection in practice, by running it end-to-end on a mixed

set of reports, where some contain residual identifiers and others are fully de-identified. We then ask human readers to assess whether any remaining PHI can be detected. Together, these components constitute GHMSCHRFT: a language-specific, practically oriented, and empirically verified framework for the trustworthy secondary use of Dutch clinical text.

The main contributions of this work are:

- A publicly available, fine-tuned Dutch clinical transformer model for PHI detection, evaluated on both internal and external test sets to assess cross-institution generalizability.
- A detailed annotation protocol with 24 granular categories for Dutch medical PHI, including span-merging and edge-case rules, provided in the Supplementary Material to support reproducible annotation.
- An open-source Dutch hiding-in-plain-sight surrogate replacement module (`https://github.com/DIAGNijmegen/dutch-med-hips`) providing type-consistent surrogate generation.
- A human reader study evaluating residual re-identification risk, providing empirical evidence of the practical protection offered by the complete end-to-end system.

Statement of Significance

Problem or Issue	Secondary use of Dutch clinical free-text is severely limited by the lack of reliable de-identification tools. Existing open-source systems were developed for English and achieve insufficient detection recall for practical deployment.
What is Already Known	Transformer-based models outperform rule-based de-identification systems when annotated training data are available. Surrogate replacement (HIPS) better preserves data utility than simple masking. No validated open-source transformer pipeline exists for Dutch clinical text.
What this Paper Adds	GHMSCHRFT is the first open-source transformer-based de-identification pipeline for Dutch clinical reports. It achieves detection recall of 0.985–0.986 across internal and external test sets, exceeding existing tools by more than 30 percentage points. A reader study confirms that the full pipeline provides strong practical protection. The model, surrogate module, and annotation guidelines are publicly released.
Who Would Benefit	Clinical researchers and hospital data governance teams seeking to enable secondary use of Dutch clinical free-text, and the broader clinical NLP community developing infrastructure for non-English health data.

2. Methods

Figure 1 provides a high-level overview of the GHMSCHRFT pipeline. The following subsections describe each component in detail.

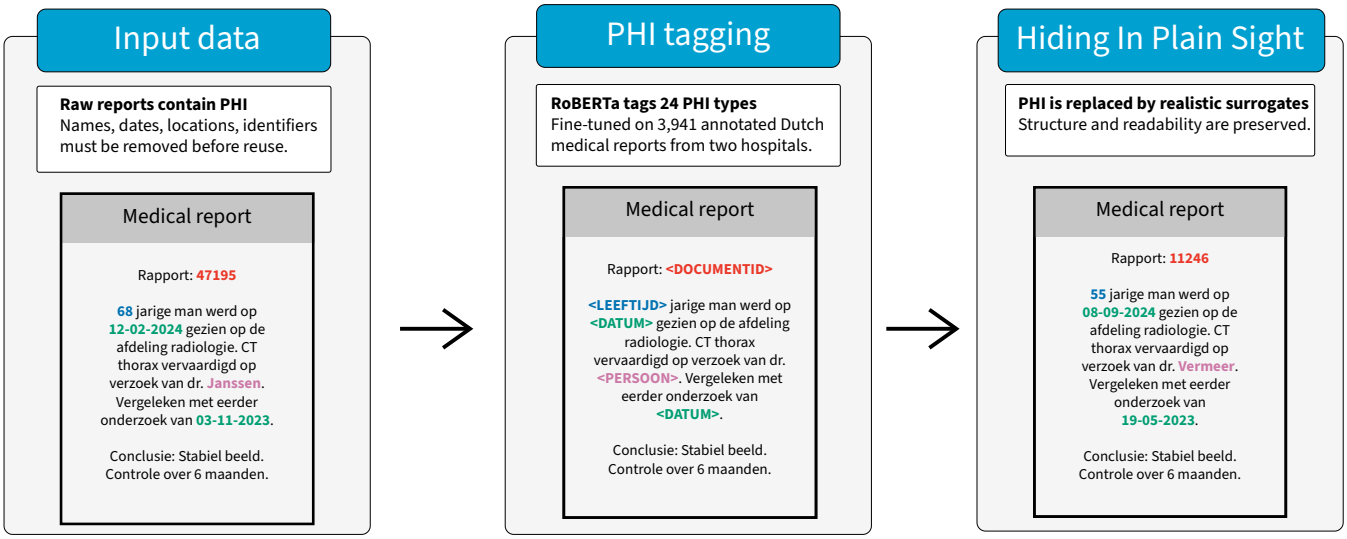


Figure 1: Overview of the GHMSCHRFT de-identification pipeline. Raw Dutch clinical reports containing PHI (left) are processed by a fine-tuned RoBERTa model that detects and tags 24 PHI categories (center). Tagged PHI is then replaced with realistic surrogates (right), preserving the report's structure and readability while rendering any missed PHI indistinguishable from the surrogates.

2.1. Data Sources and Annotation

We conducted a retrospective multicenter study to develop and evaluate GHMSCHRFT, an automated de-identification framework for Dutch clinical free-text reports. Data came from three hospitals: Radboud University Medical Center (RUMC), Ziekenhuisgroep Twente (ZGT), and Jeroen Bosch Ziekenhuis (JBZ). Development used RUMC and ZGT only: RUMC contributed manually annotated radiology and pathology reports; ZGT contributed annotated clinical and outpatient letters across eleven specialties (internal medicine, cardiology, rheumatology, pulmonology, pediatrics, surgery, allergology, gynecology, emergency medicine, gastroenterology, and urology).

JBZ contributed radiology, pathology, and emergency care reports, and was used for external validation only.

All processing was performed in secure institutional environments under GDPR and local governance policies, with ethical approval from relevant review boards. To enable GDPR-compliant cross-institutional transfer, a preliminary RUMC-trained model automatically tagged PHI at ZGT and JBZ. Annotators then corrected predictions in Doccano [19]. Prior to transfer, detected PHI spans were replaced with HIPS surrogates, preserving clinical structure while minimizing disclosure risk. Resulting annotations were stored as character-offset spans mapped to the surrogate-replaced text.

Annotation quality was ensured through a two-stage review process. Multiple readers independently annotated each report in a first pass. A subset of reports was doubly annotated to assess inter-reader agreement, quantified using character-level Cohen's kappa restricted to PHI positions. An author subsequently reviewed all annotations, resolving ambiguities and correcting errors. Annotation rules followed the PHI guidelines in Supplementary Material (Section S7).

Dataset statistics are summarised in Table 1. The internal test sets comprised 741 RUMC and 239 ZGT reports, while all 340 JBZ reports served as an external test set, excluded from all development. The remaining 3,941 reports formed the development set. Fine-grained labels were represented in Beginning-Inside-Outside (BIO) format across 24 training categories. The complete tag inventory is provided in Supplementary Table S1.

Table 1: Dataset statistics by hospital and split. RUMC: Radboud University Medical Center (Radboudumc, Nijmegen). ZGT: Ziekenhuisgroep Twente (Almelo/Hengelo). JBZ: Jeroen Bosch Ziekenhuis ('s-Hertogenbosch). Development reports were used for training and validation. HIPS-augmented duplicates and 50 synthetic reports were added during training but are not counted here. Average entities per report is shown in parentheses.

	Development		**Internal test**		**External test**	
Dataset	Reports	Entities	Reports	Entities	Reports	Entities
RUMC	2,983	8,731 (2.9)	741	2,224 (3.0)	—	—
ZGT	958	10,684 (11.2)	239	2,675 (11.2)	—	—
JBZ	—	—	—	—	340	1,818 (5.3)
Total	**3,941**	**19,415 (4.9)**	**980**	**4,899 (5.0)**	**340**	**1,818 (5.3)**

2.2. PHI Detection Model Development and Training Augmentation

PHI detection was trained by fine-tuning `dragon-roberta-large-mixed-domain` [20], a RoBERTa-large model pretrained on mixed-domain Dutch medical text, using the DRAGON baseline framework [21] with a token-classification NER head. Input sequences were left-truncated to 512 tokens during training. At inference, documents exceeding 512 tokens are processed using a strided sliding-window approach to ensure full-document coverage. Training used HuggingFace Transformers with a learning rate of 1×10^{-5}, effective batch size 8 (batch size 1, 8 gradient-accumulation steps), and gradient checkpointing; the best checkpoint was selected by validation performance.

Two augmentation strategies were applied to the training split. First, 50 fully synthetic reports were generated with `qwen3:30b-a3b-thinking-2507-q4_K_M` [22] to address class imbalance for rare identifier types (e.g., IBAN, BSN). Second, HIPS augmentation via `dutch-med-hips` [18] produced an additional surrogate-replaced version of each annotated report. The final training set combined original development reports, their HIPS-augmented counterparts, and the 50 synthetic reports.

2.3. Dutch HIPS Surrogate Module (`dutch-med-hips`)

`dutch-med-hips` [18] replaces tagged PHI with type-consistent surrogates, covering names and initials, ages, dates and times, hospital and location names, study names, contact details, and structured identifiers (BSN, IBAN, report IDs, institution-specific numbers). Surrogate generation uses a Dutch-locale fake data library, sampling predominantly Dutch names with a lower proportion of foreign names to reflect realistic patient demographics, supplemented by curated geographic lookup pools and template-driven synthesis for structured identifiers. The engine supports deterministic behavior via document-hash seeding and returns de-identified text with a structured original-to-surrogate mapping including character offsets.

At inference, GHMSCHRFT applies four sequential steps: text preprocessing, PHI detection, post-processing, and surrogate replacement. Post-processing merges adjacent same-type predictions to reduce span fragmentation from subword tokenization, after which finalized spans are replaced by the HIPS module.

2.4. Evaluation Metrics

Performance was evaluated using two primary metrics. Micro-averaged entity-level F1-score (exact span and type match, via `seqeval`/DragonEval [21]) was the principal summary measure. Detection recall served as a safety-oriented complement: a PHI entity was counted as detected when every token in the gold span received any non-`O` label, regardless of type or span extent, making false negatives under this metric a direct proxy for residual disclosure risk. Macro-averaged F1-score and per-category breakdowns are additionally reported. All headline metrics use 95% confidence intervals from document-level bootstrap resampling (5,000 iterations). As a measure of inadvertent information loss, we counted predicted PHI spans that had no overlap with any annotated gold span; these represent clinical content that the pipeline would replace with a surrogate despite not being PHI.

2.5. Baseline Comparison

GHMSCHRFT was compared against DEDUCE v3.0.6 [15] (the main open-source rule-based system for Dutch clinical text), the OpenAI Privacy Filter [14] (a general-purpose neural de-identification model), and the Radboudumc Radiology Report Anonymizer (RRA) v2.14 (an institution-specific in-house comparator). Because these systems use only partially overlapping tag inventories, comparisons were conducted across seven compiled categories with clear cross-system counterparts (Supplementary Table S7), using compiled-category detection recall on both test sets.

2.6. End-to-End Residual Risk Check

To assess residual disclosure risk in the final deployment scenario, five native Dutch-speaking readers (three clinical researchers, two radiology residents) independently reviewed 100 HIPS-processed de-identified reports in Doccano [19]. Readers marked fragments believed to be original PHI missed by the pipeline as either definite or suspected residual PHI. Readers were instructed not to flag correctly replaced surrogates and were blinded to report composition.

The 100 reports comprised the complete set of cases with at least one pipeline-missed PHI span in the combined internal test set ($n = 50$, exhausting all error cases), and an equal number of fully de-identified cases drawn from the same test set. Prior to the study, one author reviewed all 50 error cases to assess human detectability in context. Thirty-two were reclassified as effectively de-identified because the remaining fragments were undetectable

after partial surrogate replacement (e.g., a date reduced to a bare digit). This left 18 cases with 21 evaluable leaked spans; combined with the 50 fully de-identified and 32 reclassified cases, 82 of 100 reports were treated as correctly de-identified. Reader performance was quantified by counting correctly identified leaked fragments (reader annotation overlapping a gold leaked span), missed fragments, and false positives. Exact 95% confidence intervals for per-reader detection rates were computed using the Clopper–Pearson method. To quantify surrogate dilution, the precision upper bound for a perfect PHI-shape detector was computed as $L/(L+S)$, where L is the total number of evaluable leaked spans and S the total number of HIPS surrogate replacements across the 100 documents. A per-document Monte Carlo simulation (10,000 iterations, seed 42) compared each reader's observed true-positive count against the null distribution expected under uniform random sampling of the same annotation volume from the $L+S$ candidate pool.

3. Results

3.1. Inter-reader Agreement

Inter-reader agreement was high in both doubly annotated subsets. In RUMC radiology, character-level Cohen's kappa restricted to PHI positions was 0.940. In RUMC pathology, character-level Cohen's kappa restricted to PHI positions was 0.967.

3.2. Overall Internal Performance

On the combined internal test set of 980 reports containing 4,899 annotated PHI entities, the model achieved micro-averaged precision, recall, and F1-score of 0.941, 0.944, and 0.943, respectively. Detection evaluated independently of label correctness yielded precision, recall, and F1-score of 0.948, 0.950, and 0.949. Performance was strongest for structured identifier categories, with accreditation numbers, BIG registration numbers, document identifiers, and URLs all achieving F1-scores of 0.99 or higher. Lower F1-scores were observed for rarer or semantically broader categories, including company names (0.25), study names (0.55), addresses (0.55), and place names (0.62). Among the most frequent categories, F1-scores were 0.97 for dates, 0.89 for person names, 0.95 for initials, 0.90 for times, and 0.81 for hospital names.

3.3. External Validation

On the JBZ external test set of 340 reports containing 1,818 annotated PHI entities, merged-entity evaluation achieved micro-averaged precision, recall, and F1-score of 0.949, 0.952, and 0.950, respectively. Detection evaluated independently of label correctness yielded precision, recall, and F1-score of 0.981, 0.984, and 0.983. Performance patterns were broadly consistent with the internal results; per-category breakdowns are provided in Table 2.

3.4. Per-tag Performance

Table 2: Per-tag performance aggregated across the internal test sets (RUMC radiology, RUMC pathology, ZGT; $N = 980$ reports) and the external JBZ test set ($N = 340$ reports), evaluated under merged-entity scoring. Detection recall counts a gold entity as detected when every token in the annotated span receives a non-O label, irrespective of assigned entity type or predicted span extent. Precision, recall, and F1 require the exact correct entity label and span boundaries. The **Overall** row reports micro-averaged metrics; 95% bootstrap confidence intervals (5,000 iterations, document-level resampling) are shown in parentheses on the following line. Per-dataset breakdowns are provided in Supplementary Table S2.

Tag	**N**	**Det. rec.**	**Prec.**	**Rec.**	**F1**
Internal test set (RUMC & ZGT; $N = 980$)					
accreditatie nummer *(accreditation no.)*	111	1.00	1.00	1.00	1.00
adres *(address)*	11	0.91	0.55	0.55	0.55
agbnummer *(AGB no.)*	42	1.00	0.98	1.00	0.99
bedrijf *(company)*	2	0.50	0.17	0.50	0.25
bignummer *(BIG no.)*	35	1.00	1.00	1.00	1.00
bsn *(citizen no.)*	7	1.00	0.78	1.00	0.88
datum *(date)*	2406	1.00	0.98	0.97	0.97
documentid *(document ID)*	245	1.00	0.99	1.00	0.99
documentnummer *(document no.)*	252	1.00	0.99	0.99	0.99
leeftijd *(age)*	72	0.92	0.73	0.92	0.81
patientnummer *(patient no.)*	11	0.91	0.90	0.82	0.86
persoon *(person)*	592	0.96	0.91	0.87	0.89
persoonafkorting *(initials)*	429	0.98	0.94	0.95	0.95
phinummer *(other PHI no.)*	35	0.97	0.75	0.60	0.67

Continued on next page

Table 2 (continued)

Tag	N	Det. rec.	Prec.	Rec.	F1
plaats *(place)*	13	0.85	0.56	0.69	0.62
rapport id *(report ID)*	198	0.98	0.97	0.97	0.97
studie naam *(study name)*	8	1.00	0.43	0.75	0.55
telefoonnummer *(phone no.)*	28	1.00	0.74	0.89	0.81
tijd *(time)*	129	0.92	0.91	0.90	0.90
url *(URL)*	8	1.00	1.00	1.00	1.00
ziekenhuis *(hospital)*	265	0.94	0.78	0.83	0.81
Overall	**4899**	**0.983**	**0.941**	**0.944**	**0.943**
		(0.98, 0.99)	(0.93, 0.95)	(0.93, 0.95)	(0.93, 0.95)
External test set (JBZ; $N = 340$)					
accreditatie nummer *(accreditation no.)*	100	1.00	1.00	1.00	1.00
bignummer *(BIG no.)*	3	1.00	0.04	0.33	0.08
datum *(date)*	807	0.99	0.99	0.99	0.99
leeftijd *(age)*	78	0.99	0.94	0.99	0.96
patientnummer *(patient no.)*	70	1.00	0.97	0.47	0.63
persoon *(person)*	482	0.97	0.98	0.95	0.96
persoonafkorting *(initials)*	77	1.00	0.94	0.99	0.96
phinummer *(other PHI no.)*	2	1.00	0.00	0.00	0.00
plaats *(place)*	5	0.80	0.67	0.80	0.73
rapport id.c nummer *(report ID C-no.)*	41	1.00	1.00	1.00	1.00
rapport id.t nummer *(report ID T-no.)*	62	1.00	1.00	1.00	1.00
telefoonnummer *(phone no.)*	21	1.00	0.94	0.71	0.81
tijd *(time)*	58	0.97	0.97	0.97	0.97
ziekenhuis *(hospital)*	12	1.00	0.48	0.92	0.63
Overall	**1818**	**0.986**	**0.949**	**0.952**	**0.950**
		(0.98, 0.99)	(0.94, 0.96)	(0.94, 0.96)	(0.94, 0.96)

Figure 2 summarizes span-level precision and recall by PHI tag across all datasets and highlights that most frequent tag types cluster in the high-performance region, whereas lower-performing tags are generally rare and semantically broader.

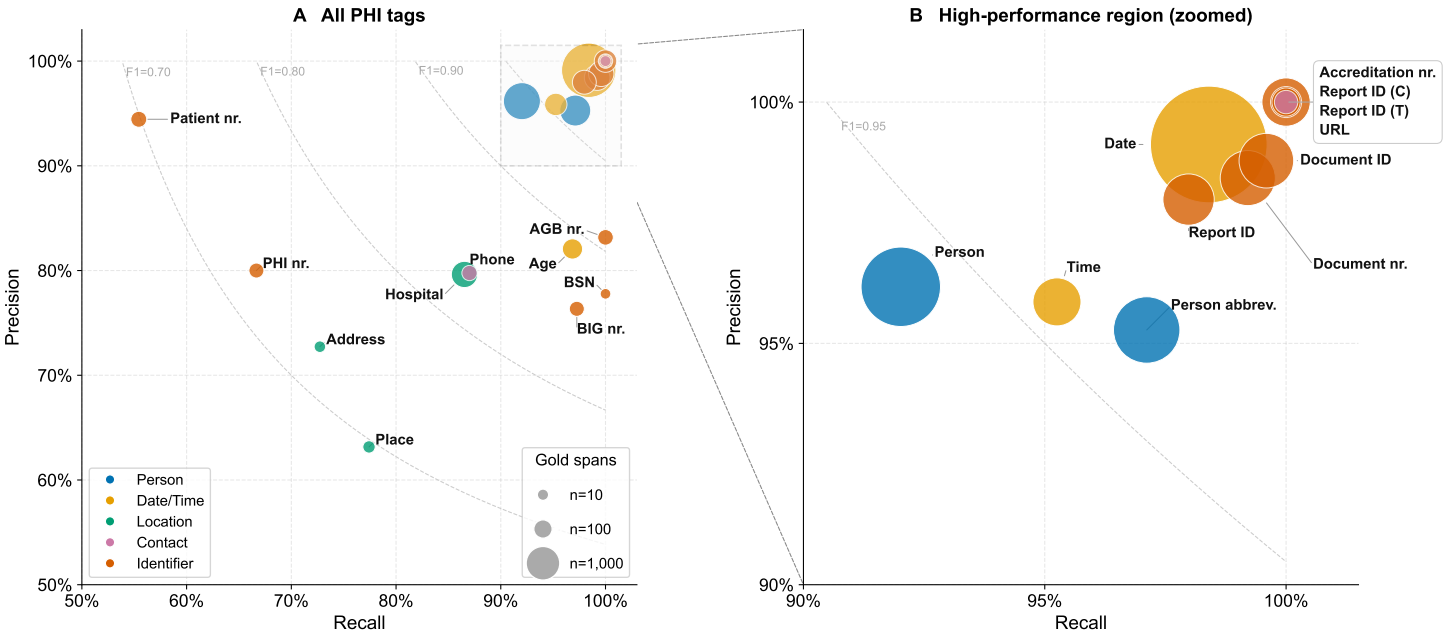


Figure 2: Span-level precision and recall per PHI tag, aggregated across all datasets. Each bubble represents one tag type; bubble area is proportional to the number of gold spans (n). Tags are coloured by PHI category. F1 iso-contours are shown as dashed lines. Both axes start at 0.5 rather than 0. (A) All tag types across the full precision–recall space. The dashed box indicates the region expanded in panel B. (B) Zoomed view of the high-performance region ($\geq$90% precision and recall). Tags that overlap at identical coordinates (100%, 100%) are listed in the inset. Span matching is exact: a predicted span is counted as correct only if its token boundaries and tag type exactly match the gold annotation.

3.5. Comparison with Rule-based Baselines

GHMSCHRFT achieved an overall compiled-category detection recall of 0.985 (95% CI 0.98–0.99) on the combined internal test data, compared with 0.549 (0.53–0.57) for DEDUCE, 0.386 (0.36–0.41) for RRA, and 0.339 (0.31–0.36) for the OpenAI Privacy Filter. On the external JBZ test set, GHMSCHRFT retained strong performance with an overall recall of 0.986 (0.98–0.99) and substantially outperformed DEDUCE (0.682, 0.65–0.72), RRA (0.295, 0.27–0.32), and the OpenAI Privacy Filter (0.498, 0.46–0.54). Per-category and per-dataset breakdowns are provided in Supplementary Table S5.

Across the internal datasets, GHMSCHRFT showed the largest gains for person names, dates, hospital names, and compiled identifier categories. DEDUCE remained competitive for age expressions and some date-heavy subsets, whereas RRA performed best on the in-domain radiology subsets but generalized poorly to ZGT and JBZ. The OpenAI Privacy Filter showed moderate recall for person names and dates in radiology subsets but near-zero recall for hospital names and age expressions across all datasets, likely reflecting a domain and language mismatch with Dutch clinical text. Figure 3 visualizes these category-level and dataset-level differences.

3.6. False-Positive Rate and Information Loss

Of 6,734 predicted PHI spans across both test sets, 6,353 (94.3%) were exact true positives, 271 (4.0%) overlapped a gold span but with incorrect boundaries or tag type, and 110 (1.6%) had no overlap with any gold span. These 110 pure false positives comprised 132 tokens across 1,320 reports (0.08 per report), representing clinical content that would be replaced with a surrogate in a deployed system. Full per-tag results are provided in Supplementary S3.

3.7. End-to-End Residual Risk Assessment

After applying the exclusions described in the Methods, 18 of the 100 presented cases contained evaluable residual PHI, comprising 21 leaked PHI spans in total. Per-reader detection rates with exact 95% confidence intervals are reported in Table 3. No reader identified more than three spans: DS identified 3 (14.3%; 95% CI 3.0–36.3%), RW and HK each identified 2 (9.5%; 1.2–30.4%), SB identified 1 (4.8%; 0.1–23.8%), and MB identified none (0.0%; 0.0–16.1%). Confidence intervals are wide due to the small number of evaluable spans. Collectively, five of the 21 leaked spans were identified by at least

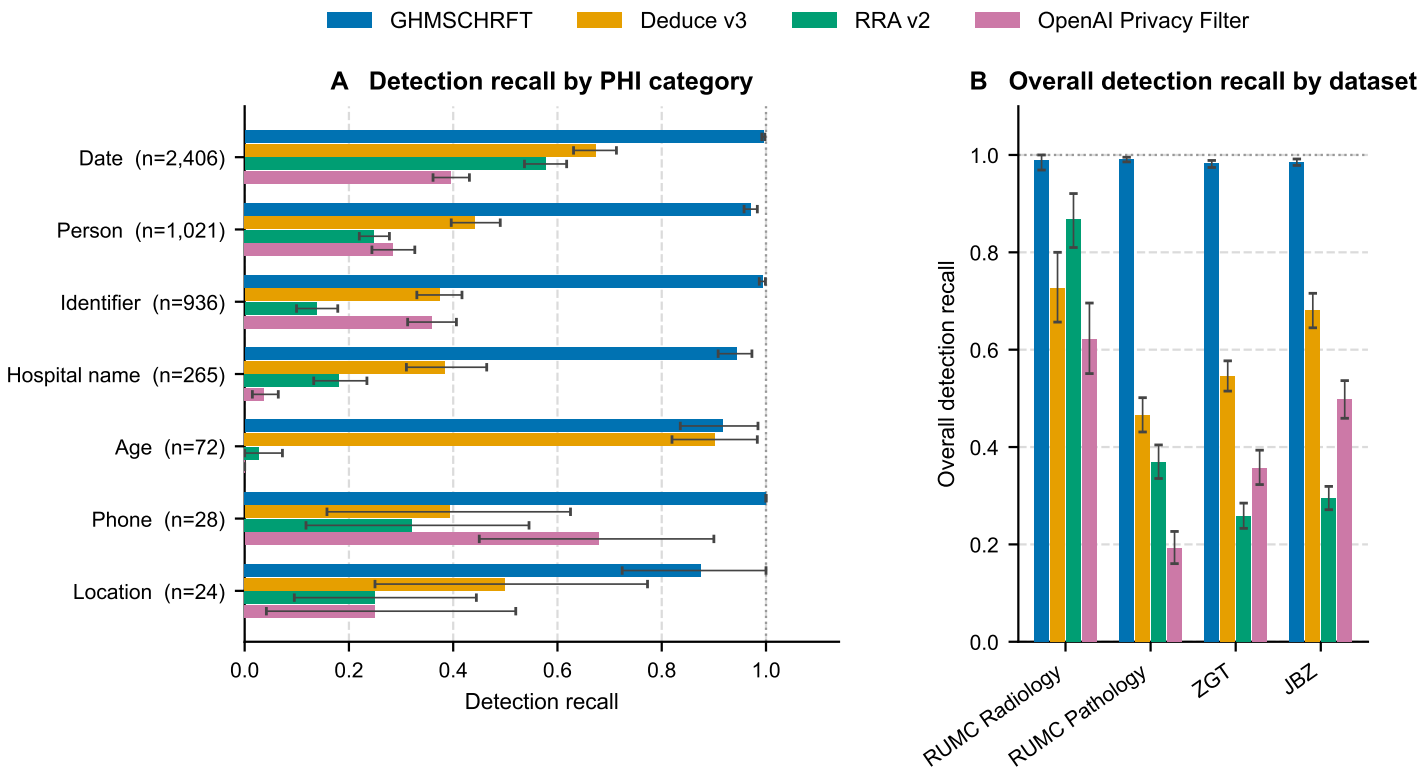


Figure 3: Detection recall per PHI category and per dataset for GHMSCHRFT and three comparator systems. A span is counted as detected if every token in the gold span receives a non-O prediction, irrespective of assigned label. (A) Per-category detection recall on the combined internal test set; n denotes the number of gold spans per category. Categories are ordered by frequency. (B) Overall detection recall per system across four hospital datasets. GHMSCHRFT was trained on RUMC and ZGT data; JBZ is an unseen external institution. DEDUCE v3 and RRA v2 are rule-based Dutch clinical de-identification systems. The OpenAI Privacy Filter is a 1.5B sparse mixture-of-experts model trained on general, primarily English text. Detection recall is reported rather than F1 to reflect the safety-critical requirement that no PHI is missed. Error bars denote 95% bootstrap confidence intervals (5,000 iterations, document-level resampling).

one reader: a partially-leaked surname, an age fragment, a hospital panel abbreviation, and a first name appearing twice in the same document (each occurrence counted as a separate span). The remaining 16 spans went undetected by every reader.

Readers generated substantially more annotations that did not correspond to any leaked PHI (false positives). RW made 57 annotations in total, of which 55 were false positives; DS made 36, of which 34 were false positives; MB made 34, all false positives; SB made 33, of which 32 were false positives; and HK made 25, of which 24 were false positives. Across readers, 0–28% of flagged error cases included at least one 'definite PHI' annotation. The remainder were labeled 'suspected PHI'. False-positive annotations arose under both confidence levels: several readers definitively flagged HIPS-generated surrogate names as leaked PHI while overlooking actual residual content.

A surrogate dilution analysis quantified the structural precision limit. The 100 reader-study documents contained $S = 1{,}039$ HIPS surrogate replacements alongside $L = 21$ evaluable leaked spans, yielding a combined PHI-shaped pool of 1,060 candidates. The theoretical precision ceiling for a perfect PHI-shape detector on this dataset is $L/(L + S) = 2.0\%$. All observed reader precisions (0%–8.3%) fell within the 95% interval of the per-document Monte Carlo simulation, confirming that no reader performed significantly above the level expected from random selection within this pool.

Table 3: Span-level results per reader in the end-to-end residual risk assessment. The 18 evaluable error cases contained 21 leaked PHI spans in total. Detection rate is the fraction of leaked PHI spans overlapped by at least one reader annotation; 95% confidence intervals are exact Clopper–Pearson intervals. False positives (FP) are reader annotations not overlapping any gold leaked fragment. Spans found may exceed TP reader annotations when a single reader annotation simultaneously overlapped two adjacent leaked spans.

Reader	**Role**	**Spans found**	**Detection rate (95% CI)**	**FP**
RW	Clinical researcher	2/21	9.5% (1.2–30.4%)	55
SB	Clinical researcher	1/21	4.8% (0.1–23.8%)	32
MB	Clinical researcher	0/21	0.0% (0.0–16.1%)	34
HK	Radiology resident	2/21	9.5% (1.2–30.4%)	24
DS	Radiology resident	3/21	14.3% (3.0–36.3%)	34

4. Discussion

GHMSCHRFT provides an end-to-end Dutch de-identification pipeline with high detection recall, external generalization, and practical residual-risk evaluation, addressing a gap that has persisted since rule-based systems became the de facto standard. Near-identical performance on the internal and external JBZ test sets on radiology, pathology, and emergency care reports suggests the model has learned generalizable representations of Dutch clinical PHI rather than institution-specific formatting conventions for these report types. ZGT letters were part of internal development, so the evidence for generalization in that setting is within-institution only. This likely reflects the relevance of the pretraining domain: Bosma et al. [21] demonstrated that Dutch medical pretraining consistently improves clinical NLP performance, and the present results extend that finding to de-identification.

GHMSCHRFT's recall advantage over the rule-based baselines is best illustrated by RRA. Built specifically for RUMC radiology reports, RRA achieved high recall on that subset but dropped markedly on ZGT and JBZ, and notably, even on RUMC pathology reports, despite originating from the same institution. This generalization failure is typical of rule-based systems, which struggle whenever reporting templates or documentation conventions differ from those they were built for. DEDUCE showed a similar, though less pronounced, shortfall. It was, however, evaluated without institution-specific tuning. Manual fine-tuning per target institute could improve its performance.

OpenAI Privacy Filter's poor performance has a different explanation. As a neural model trained on general, primarily English text, its near-zero recall for hospital names and age expressions reflects a mismatch with Dutch clinical text rather than an architectural limitation. To probe this further, we additionally evaluated a version fine-tuned on English medical text [23]. Full results are provided in Supplementary S5. This fine-tuned variant did not outperform the out-of-the-box model on our Dutch data. As its fine-tuning remained in English, it addressed neither the Dutch language itself nor the Netherlands-specific institutional and identifier conventions that distinguish Dutch clinical documentation.

At a more granular level, per-tag performance was strongest for structurally well-defined identifier types. Lower F1-scores for company names (0.25), study names (0.55), addresses (0.55), and place names (0.62) are primarily attributable to low training support rather than inherent difficulty, as the uniformly high detection recall shows the model generally identified these

spans as PHI. Errors were more often misclassification or boundary mistakes than outright missed detections. For addresses, errors were almost exclusively boundary errors. For example, a full street address would be correctly detected with only a trailing province abbreviation excluded from the predicted span.

Two related anomalies appeared in the external test set: patient numbers achieved perfect detection recall but an F1 of only 0.63, while BIG registration numbers showed precision of just 0.04 despite only three gold instances. Manual inspection revealed a shared cause. The model consistently detected patient number spans but misclassified them as BIG numbers, because JBZ uses a numeric patient identifier format superficially resembling BIG numbers not seen during training. This is a cross-institution format mismatch, not a detection failure.

A separate, more critical failure mode appeared primarily in clinical letters, where patients are occasionally referred to by first name only. This pattern was absent from radiology and pathology reports that dominated training. The model had limited exposure to this and did not reliably recognize isolated first names as PHI in clinical correspondence. Future work will address this by including additional annotated examples.

The false-positive analysis (Supplementary S3) indicates that the pipeline accidentally destroys very little useful clinical information. Only 110 of 6,734 predicted spans (1.6%) had no overlap with any gold PHI annotation, affecting 132 tokens across 1,320 reports (0.08 per case). Hospital names (ZIEKENHUIS, 27 cases) and age expressions (LEEFTIJD, 23 cases) were the main contributors, reflecting surface ambiguity with institution-like abbreviations and clinical numeric values respectively.

Detection-only evaluation is incomplete without assessing whether residual PHI remains recognizable in the final output. The reader study provides empirical evidence for the practical protection offered by the complete pipeline. After HIPS processing, readers collectively identified only five of 21 evaluable leaked spans, with no individual reader identifying more than three (DS: 14.3%; 95% CI 3.0–36.3%). Simultaneously, readers generated 24–55 false positives each, consistently flagging correctly replaced surrogates as suspicious while overlooking genuine residual PHI.

The surrogate dilution analysis provides a mechanistic interpretation of this finding. The 100 reviewed documents contained $S = 1{,}039$ HIPS surrogate replacements alongside $L = 21$ evaluable leaked spans, giving a theoretical precision ceiling of $L/(L+S) = 2.0\%$ for any reader who cannot distinguish surrogates from genuine leaks by surface form alone. All observed

reader precisions (0%–8.3%) were consistent with random selection from this pool under per-document Monte Carlo simulation. Protection is therefore not solely a function of detection recall, as the large volume of realistic surrogates structurally dilutes residual leaks to the point where they are statistically indistinguishable from correctly replaced content.

This work also has limitations that point to important directions for future research. External validation was limited to pathology, radiology, and emergency care letter types from JBZ. Equivalent performance on outpatient and clinical letters from unseen institutions cannot be assumed. Rare PHI categories remain challenging despite augmentation, and the synthetic reports introduce a dependency on the quality of the generative model. Finally, this study does not provide formal re-identification risk quantification. Detection recall and reader performance are proxies, and their relationship to actual re-identification probability under a realistic adversary model remains an open question.

Beyond the Dutch setting, the bootstrapping approach using a preliminary model to enable GDPR-compliant cross-institutional data transfer through automated tagging and surrogate replacement may offer a practical model for other language communities assembling multicenter corpora under similar regulatory constraints. The results also suggest that HIPS-based pipelines warrant broader adoption. The residual risk profile of a well-calibrated surrogate replacement system may be substantially more favorable than detection recall figures alone imply, as even imperfect detection can provide strong practical protection.

5. Conclusion

We presented GHMSCHRFT, an end-to-end de-identification pipeline for Dutch clinical free-text combining a fine-tuned transformer model for PHI detection with a Dutch-specific HIPS surrogate replacement module. The pipeline achieved strong, consistent performance across internal and external test sets, demonstrating cross-institutional generalizability for radiology, pathology, and emergency care reports, and substantially outperforming rule-based baselines. The reader study further showed that after surrogate replacement, residual PHI was difficult for domain-familiar readers to distinguish from surrounding HIPS-generated content. Together, high detection recall and realistic surrogate replacement make the pipeline robust to the imperfect detection that is inevitable in practice. By releasing the detection

model, Dutch HIPS module, and annotation guidelines, this work provides both a practical tool for Dutch clinical text de-identification and a reusable framework for other clinical NLP settings operating under similar regulatory constraints.

Declaration of Competing Interest

J.S. Bosma is employed by kaiko.ai. B. van Ginneken is a shareholder of Thirona BV and Plain Medical. M. Prokop reports grants paid to his institution from the Dutch Cancer Society (KWF), the European Commission, and the Dutch government (ZonMW); honoraria and royalties paid to his institution from Canon Medical Systems and Bracco SA; patents held by Radboud University Medical Center, including one jointly with Canon Medical Systems; an advisory board role at Fraunhofer MEVIS; leadership roles at the European Society of Radiology and the Dutch Radiological Society; and is a co-founder and shareholder of Plain Medical. All other authors declare that they have no competing interests.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Claude (Anthropic) for writing and coding assistance. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

Data Availability

The clinical data used in this study are not publicly available due to patient privacy constraints and applicable data protection regulations.

Code Availability

The code for running the GHMSCHRFT pipeline is publicly available at `https://github.com/DIAGNijmegen/GHMSCHRFT`. The code for training the PHI detection model and running all evaluation for this manuscript is available at `https://github.com/DIAGNijmegen/GHMSCHRFT-training`.

CRediT authorship contribution statement

Luc Builtjes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Joeran S. Bosma:** Conceptualization, Methodology, Software, Data curation, Writing – review & editing. **Judith Lefkes:** Validation, Data curation, Writing – review & editing. **Anouk Veldhuis:** Data curation, Writing – review & editing. **Jeroen Geerdink:** Data curation, Writing – review & editing. **Tijmen de Haas:** Data curation, Writing – review & editing. **Fabian R. Hoitsma:** Data curation. **Rianne A. Weber:** Investigation, Writing – review & editing. **Sarah de Boer:** Investigation, Writing – review & editing. **Myrthe Buser:** Investigation, Writing – review & editing. **Harry J. Kendall:** Investigation, Writing – review & editing. **Delano Sanches:** Investigation, Writing – review & editing. **Bram van Ginneken:** Supervision, Writing – review & editing. **Henkjan Huisman:** Supervision, Writing – review & editing. **Ewoud J. Smit:** Supervision, Writing – review & editing. **Mathias Prokop:** Supervision, Writing – review & editing. **Alessa Hering:** Conceptualization, Supervision, Writing – review & editing.

References

[1] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, T. Botsis, Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review, Journal of Biomedical Informatics 73 (2017) 14–29. doi:10.1016/j.jbi.2017.07.012.

[2] B. Percha, Modern clinical text mining: a guide and review, Annual Review of Biomedical Data Science 4 (1) (2021) 165–187. doi:10.1146/annurev-biodatasci-030421-030931.

[3] T. M. Seinen, E. A. Fridgeirsson, S. Ioannou, D. Jeannetot, L. H. John, J. A. Kors, A. F. Markus, V. Pera, A. Rekkas, R. D. Williams, et al., Use of unstructured text in prognostic clinical prediction models: a systematic review, Journal of the American Medical Informatics Association 29 (7) (2022) 1292–1302. doi:10.1093/jamia/ocac058.

[4] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning

with electronic health records, npj Digital Medicine 1 (1) (2018) 18. doi:10.1038/s41746-018-0029-1.

[5] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council, https://eur-lex.europa.eu/eli/reg/2016/679/oj, general Data Protection Regulation (2016).

[6] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, M. H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Medical Research Methodology 10 (1) (2010) 70. doi:10.1186/1471-2288-10-70.

[7] A. E. W. Johnson, L. Bulgarelli, T. J. Pollard, Deidentification of free-text medical records using pre-trained bidirectional transformers, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 214–221. doi:10.1145/3368555.3384455.

[8] A. Kovačević, B. Bašaragin, N. Milošević, G. Nenadić, De-identification of clinical free text using natural language processing: a systematic review of current approaches, Artificial Intelligence in Medicine 151 (2024) 102845. doi:10.1016/j.artmed.2024.102845.

[9] I. C. Wiest, M.-E. Leßmann, F. Wolf, D. Ferber, M. v. Treeck, J. Zhu, M. P. Ebert, C. B. Westphalen, M. Wermke, J. N. Kather, Deidentifying medical documents with local, privacy-preserving large language models: the LLM-anonymizer, NEJM AI 2 (4) (2025) AIdbp2400537. doi:10.1056/AIdbp2400537.

[10] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, et al., DeID-GPT: zero-shot medical text de-identification by GPT-4, arXiv preprint arXiv:2303.11032 (2023). doi:10.48550/arXiv.2303.11032.

[11] K. Aghakasiri, N. Zambare, J. Thai, C. Ye, M. Mehta, J. R. Mitchell, M. Abdalla, Not what the doctor ordered: surveying LLM-based de-identification and quantifying clinical information loss, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 32175–32191. doi:10.18653/v1/2025.emnlp-main.1637.

[12] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al., Extracting training data from large language models, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2633–2650.

[13] OpenMed, PII and de-identification models, https://huggingface.co/collections/OpenMed/pii-and-de-identification (2023).

[14] OpenAI, OpenAI privacy filter, https://huggingface.co/openai/privacy-filter (2026).

[15] V. Menger, F. Scheepers, L. M. van Wijk, M. Spruit, DEDUCE: a pattern matching method for automatic de-identification of Dutch medical text, Telematics and Informatics 35 (4) (2018) 727–736. doi:10.1016/j.tele.2017.08.002.

[16] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, L. Hirschman, Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text, Journal of the American Medical Informatics Association 20 (2) (2013) 342–348. doi:10.1136/amiajnl-2012-001034.

[17] D. S. Carrell, B. A. Malin, D. J. Cronkite, J. S. Aberdeen, C. Clark, M. Li, D. Bastakoty, S. Nyemba, L. Hirschman, Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers, Journal of the American Medical Informatics Association 27 (9) (2020) 1374–1382. doi:10.1093/jamia/ocaa095.

[18] L. Builtjes, J. S. Bosma, A. Hering, DIAGNijmegen/dutch-med-hips: v1.0.1-zenodo (2026). doi:10.5281/zenodo.18415297.

[19] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: text annotation tool for humans (2018).
URL https://github.com/doccano/doccano

[20] J. Bosma, dragon-roberta-large-mixed-domain (revision 711c1d3) (2024). doi:10.57967/hf/2170.
URL https://huggingface.co/joeranbosma/dragon-roberta-large-mixed-domain

[21] J. S. Bosma, K. Dercksen, L. Builtjes, R. André, C. Roest, S. J. Fransen, C. R. Noordman, M. Navarro-Padilla, J. Lefkes, N. Alves, et al., The DRAGON benchmark for clinical NLP, npj Digital Medicine 8 (1) (2025) 289. doi:10.1038/s41746-025-01626-x.

[22] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). doi:10.48550/arXiv.2505.09388.

[23] OpenMed, OpenMed/privacy-filter-nemotron: fine-grained pii extraction with 55 categories, https://huggingface.co/OpenMed/privacy-filter-nemotron (2026).